

Titanic Survivor Analysis

Samrudha Medhekar | Data Analytics in R | 30/07/2018

# Executive

Using the data set, one can observe whether 891 passengers of the Titanic survived or perished, and the relationship with several variables, including Age, Sex, passenger class, if they had family on-board the ship, their ticket number, how much they paid for their ticket, where they boarded the ship, and their cabin’s location.

## INtroduction

The data analysis on the Titanic dataset was performed by Samrudha Medhekar. We explore the data to:

Understand the data

Summarize the data

Clean and Prune the data

Understand relationships between atrributes

Think about and source other data which maybe useful in answering the question

Get a preliminary feel for the types of models we think would best fit the data

## BACKGROUND

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

## Purposes

The purpose of analyzing the data of RMS Titanic is to find the relation between the number of survivors based on various factors, like gender, age, port of embarkation, class.

## Limitations

1. The data analysis provides the results of only the training set. For a complete analysis, both data sets need to be taken into account.

## methods

The following methods and techniques were used in the analysis of the dataset:-

1. As.factor function – To convert columns with categorical data into factor columns.
2. Summary and str function – To perform exploratory data analysis.
3. Na.omit() function – To perform data cleaning by removing NAs.
4. Ifelse statement – To group the age variable into adult, child and senior categories.
5. Barplot() function – To plot bar plots of different features.
6. Anova one way analysis – For validating the plots.

## Sample:-

The analysis was carried out by Samrudha Medhekar under the guidance of Neeraj Sir for Suven Consultants.

## Instrumentation:-

The following tools and libraries were used to perform the analysis:-

1. R Studio :- For performing the analysis in R language.
2. MS Excel :- For fetching the dataset.

## RESULTS:-

Total samples are 891 of the actual number of passengers on board the Titanic (2,224).

Pclass=3 had most passengers, however most did not survive.

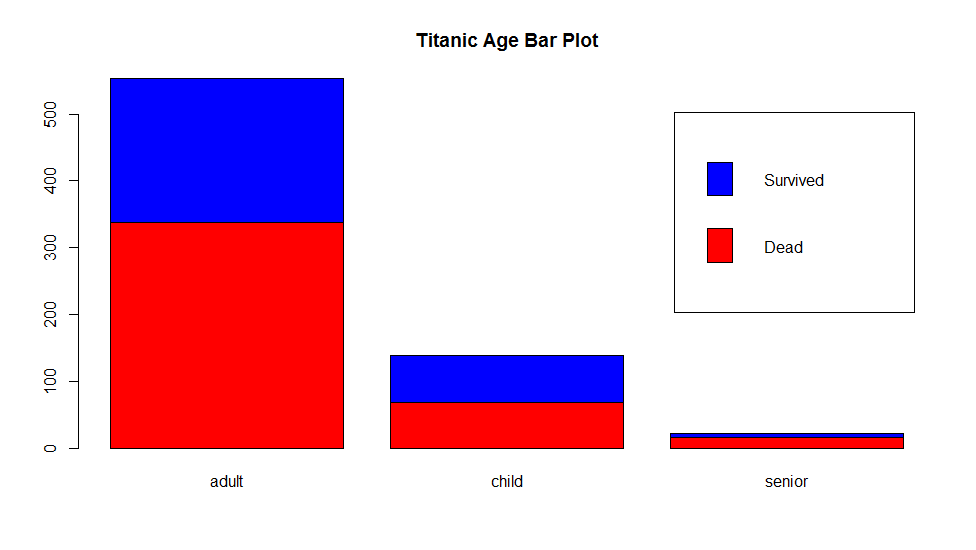
Female passengers had much better survival rate than males except in Embarked=C where males had higher survival rate.

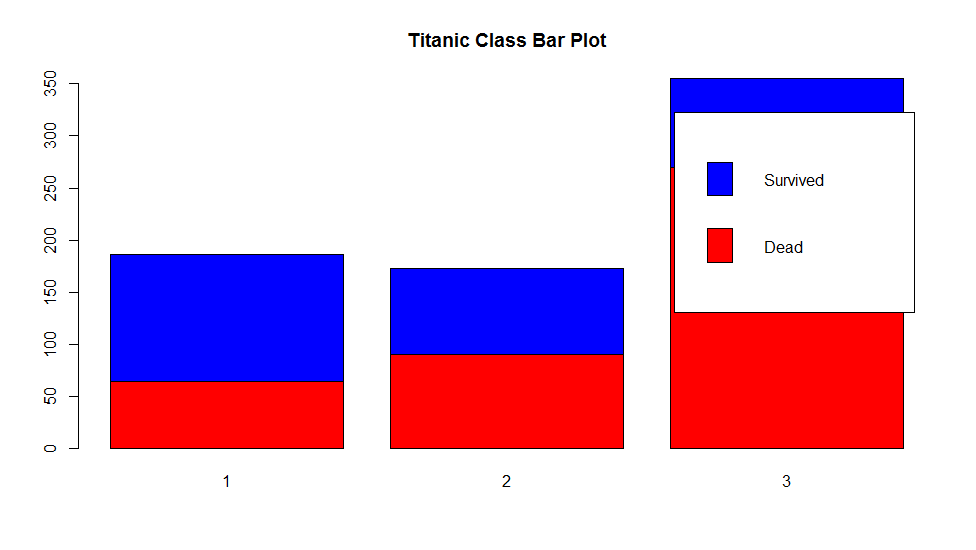
Infants (Age <=4) had high survival rate.

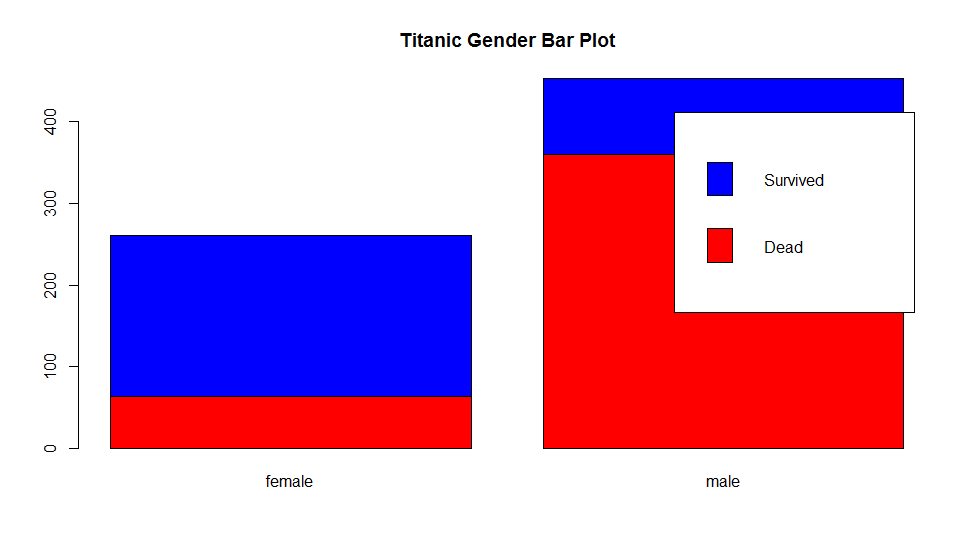
Oldest passengers (Age = 80) survived.

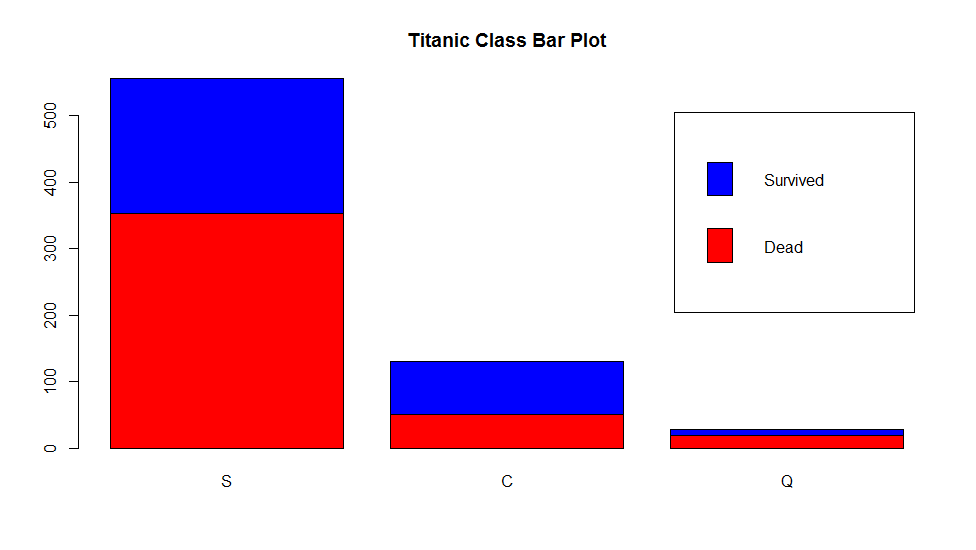
Large number of 15-25 year olds did not survive.

Most passengers are in 15-35 age range.









## REcommendationS:-

The dataset analysis was carried out on training set only, so it is recommended to carry out similar analysis on the testing set. Predictive algorithms like Multiple Regression or Decision Tree can be applied on both the datasets to predict survival rate.

This simple analysis confirms our assumptions as decisions for subsequent workflow stages.

We should consider Age in our model training.

Complete the Age feature for null values .

We should band age groups .

## Summary:-

From the analysis, we can conclude that there is a high correlation between two factors:-

1. The passenger’s gender and their class.
2. The survivors and their port of embarkment.

Females had a higher survival rate than males. Children had a nearly 50-50% survival chances, while senior citizens had the lowest chance of survival. People who boarded the first class were far more likely to survive than those who had boarded second and third class. People who embarked from the port of Queenstown had a higher survival rate (65%) than those who embarked from Cherboug or Southampton.

## REferenceS:-

<https://www.kaggle.com/monfilier/titanic-survival-prediction>

<https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic>

<https://www.kaggle.com/startupsci/titanic-data-science-solutions>

<https://www.kaggle.com/c/titanic/discussion/10151>

## 